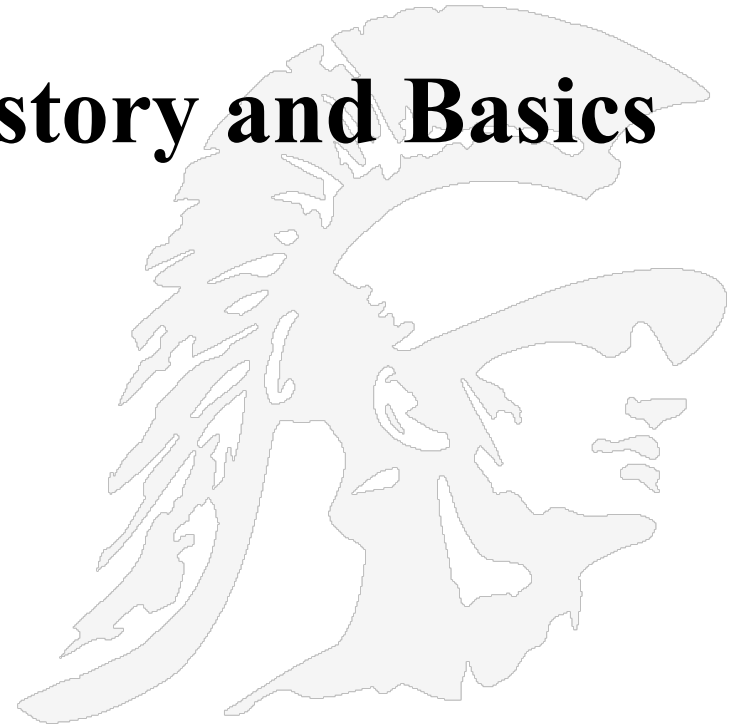




USC Viterbi
School of Engineering

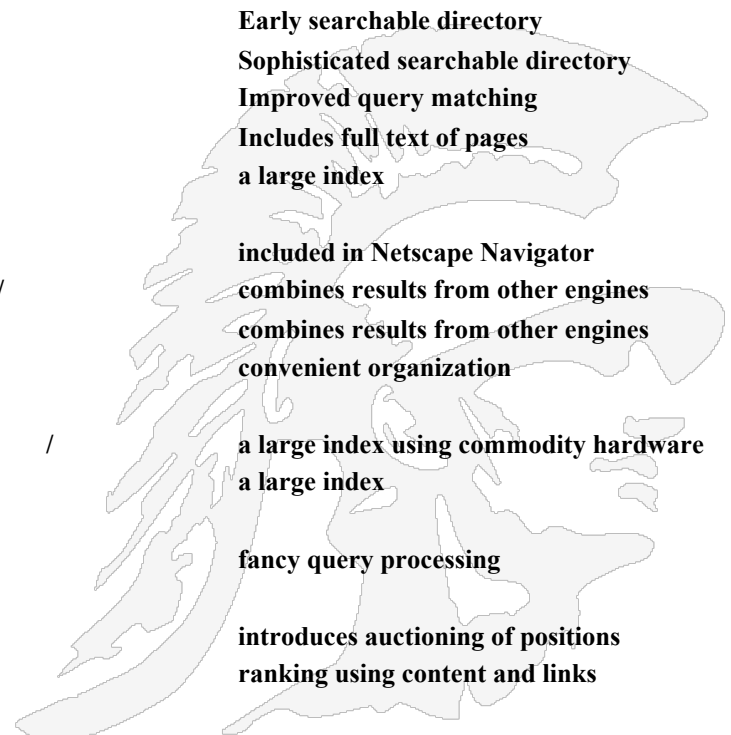
Search Engine History and Basics





A Brief Chronology of Search Engines

- 1991
 - Gopher, Archie, Veronica early search engines, non-web
- 1993
 - Wanderer,
 - ALIWeb
 - Excite powerful indexing
- 1994
 - Galaxy Early searchable directory
 - Yahoo Sophisticated searchable directory
 - Lycos Improved query matching
 - WebCrawler Includes full text of pages
 - Alta Vista a large index
- 1995
 - Infoseek included in Netscape Navigator
 - Metacrawler combines results from other engines
 - SavvySearch combines results from other engines
 - LookSmart convenient organization
- 1996
 - Inktomi a large index using commodity hardware
 - HotBot, a large index
- 1997
 - AskJeeves fancy query processing
- 1998
 - Goto introduces auctioning of positions
 - Google ranking using content and links



- Today there are hundreds of search engines, many are specialized
- See <http://www.searchenginehistory.com/> Search Engine History
- A very long web page describing the history of search engines since the 1940s to 2011-2022

Archie, Veronica, Gopher

- **By late 1980's many files were available by anonymous FTP.**
- **In 1990, Alan Emtage, P. Deutsch, et al of McGill Univ. developed Archie (short for “archives”)**
 - Assembled lists of files available on many FTP servers.
 - Allowed regex search of these file names.
- **In 1993, Veronica and Jughead were developed to search names of text files available through Gopher servers**
 - The **Gopher protocol** is a TCP/IP application layer protocol designed for distributing, searching, and retrieving documents over the Internet. Strongly oriented towards a menu-document design
 - The Gopher ecosystem is often regarded as the effective predecessor of the World Wide Web



- Excite came from the project Architext, which was started in February, 1993 by six Stanford undergrad students.
 - They had the idea of using statistical analysis of word relationships to make searching more efficient.
 - They were soon funded, and in mid 1993 they released copies of their search software for use on web sites.
- Later developments
 - Excite was bought by a broadband provider named @Home in January, 1999 for \$6.5 billion, and was named Excite@Home. In October, 2001 Excite@Home filed for bankruptcy. InfoSpace bought Excite from bankruptcy court for \$10 million
 - www.excite.com still exists as a portal

World Wide Web Wanderer

- In June 1993 Matthew Gray while at MIT introduced the World Wide Web Wanderer.
 - Initial goal was to measure the growth of the web by counting active web servers. He soon upgraded the software to capture actual URL's. His database became known as the Wandex.
- The World Wide Web Wanderer was a Perl-based web crawler that was first deployed in June 1993
- Matthew Gray now works for Google.
- While the Wanderer was probably the first web robot, and, with its index, clearly had the potential to become a general-purpose WWW search engine it never went that far
- The Wanderer charted the growth of the web until late 1995.



- In November of 1993 Martijn Koster created “Archie-Like Indexing of the Web”, or ALIWEB in response to the Wanderer.
 - Some consider it to be the first Web search engine
- ALIWEB crawled meta information and allowed users to submit their pages they wanted indexed with their own page description.
- This meant it needed no bot to collect data and was not using excessive bandwidth.
- One downside of ALIWEB was that people did not know how to submit their site

- AltaVista debut online came during December, 1995. AltaVista brought many important features to the web scene.
 - They were the first to allow natural language queries
 - They offered advanced searching techniques
 - They allowed users to add or delete their own URL within 24 hours.
 - They even allowed inbound link checking. AltaVista also provided numerous search tips and advanced search features.
- Later developments
 - On February 18, 2003, Overture signed a letter of intent to buy AltaVista for \$80 million in stock and \$60 million cash. After Yahoo! bought out Overture they rolled some of the AltaVista technology into Yahoo! Search, and occasionally used AltaVista as a testing platform.



- Lycos was designed at Carnegie Mellon University around July of 1994. Michael Loren Mauldin was responsible for this search engine and was the chief scientist at Lycos Inc in the early years.
- On July 20, 1994, Lycos went public with a catalog of 54,000 documents.
 - In addition to providing ranked relevance retrieval, Lycos provided prefix matching and word proximity bonuses.
 - Lycos' main difference was the sheer size of its catalog: by August 1994, Lycos had identified 394,000 documents; by January 1995, the catalog had reached 1.5 million documents; and by November 1996, Lycos had indexed over 60 million documents -- more than any other Web search engine.
- In October 1994, Lycos ranked first on Netscape's list of search engines
- Lycos has gone through a series of owners, and it still exists as www.lycos.com



- Infoseek also started out in 1994, founded by Steve Kirsch
- In December 1995 they convinced Netscape to use them as their default search engine, which gave them major exposure.
- One popular feature of Infoseek was allowing webmasters to submit a page to the search index in real time, which was a search spammer's paradise
- They were the first search engine to sell advertising on a CPM (Cost per Thousand) impressions basis
- Infoseek was bought by Walt Disney Company in 1998



- In 1994, two Stanford Ph.D. students David Filo and Jerry Yang posted web pages with links on them, organized into a topical hierarchy.
- As the number of links began to grow, they developed a hierarchical listing. As the pages become more popular, they developed a way to search through all of the links.
- Early on all the links on the pages were updated manually rather than automatically by spider or robot and the search feature searched only those links
- Yahoo home page acted as a portal with Email, Finance, and Groups being very successful; however after 2000 usage declined
- After many years of decline Yahoo was purchased by Verizon in 2017 for \$4.48 billion, and it lives on

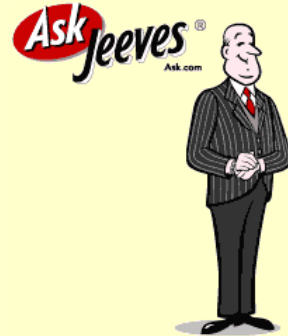


- Looksmart was founded in 1995 in Australia. They competed with the Yahoo! Directory by frequently increasing their inclusion rates
- Later developments
 - In 2002 Looksmart transitioned into a pay per click provider, which charged listed sites a flat fee per click. They syndicated those paid listings to some major portals like MSN.
 - The problem was that Looksmart became too dependant on MSN, and in 2003, when Microsoft announced they were dumping Looksmart that basically killed their business model.
 - In March of 2002, Looksmart bought a search engine by the name of WiseNut, but it never gained traction
- See <https://en.wikipedia.org/wiki/LookSmart>



- The Inktomi Corporation came about on May 20, 1996 with its search engine Hotbot. Two Cal Berkeley cohorts created Inktomi from the improved technology gained from their research
- Later developments
 - In October of 2001 Inktomi accidentally allowed the public to access their database of spam sites, which listed over 1 million URLs at that time.
 - Inktomi pioneered *the paid inclusion model* in which a website pays a fee to the search engine that guarantees the site will be displayed when certain search terms are entered
 - The model was nowhere near as efficient as the pay-per-click auction model developed by Overture. Licensing their search results also was not profitable enough to pay for their scaling costs. They failed to develop a profitable business model, and sold out to Yahoo! for approximately \$235 million, or \$1.65 a share, in December of 2003.

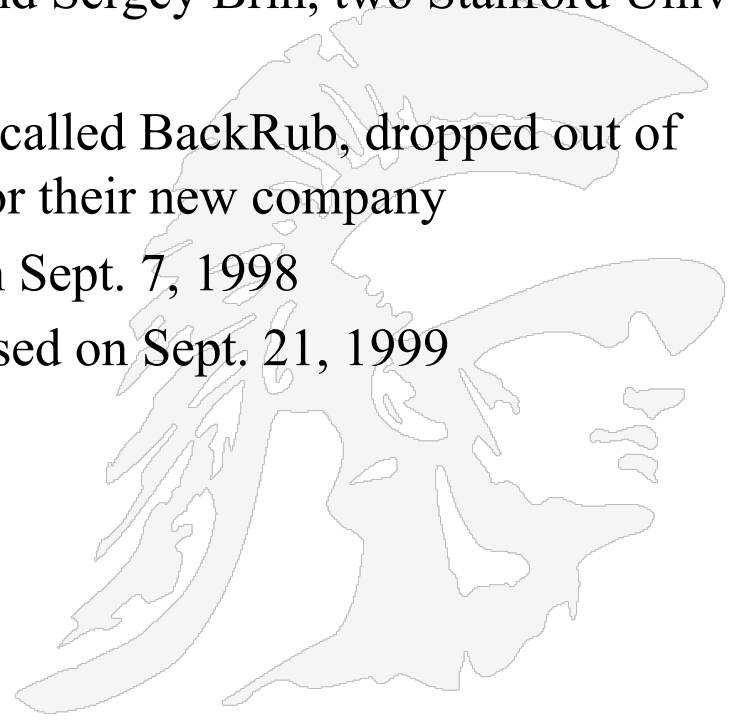
*<http://searchenginewatch.com/article/2066745/Inktomi-Spam-Database-Left-Open-To-Public>



- In April of 1997 Ask Jeeves was launched as a natural language search engine.
 - Ask Jeeves used human editors to try to match search queries.
 - Ask was powered by DirectHit for a while, which aimed to rank results based on their popularity, but that technology proved too easy to spam.
 - In 2000 the Teoma search engine was released, which uses clustering to organize sites by Subject Specific Popularity, which is another way of saying they tried to find local web communities. In 2001 Ask Jeeves bought Teoma to replace the DirectHit search technology.
 - On March 21, 2005 Barry Diller's IAC agreed to acquire Ask Jeeves for 1.85 billion dollars. IAC owns many popular websites like Match.com, Ticketmaster.com, and Citysearch.com, and is promoting Ask across their other properties.
 - In 2006 Ask Jeeves was renamed to Ask.



- Google is a play on the word Googol, coined by Milton Sirotta; it refers to a 1 followed by 100 zeros, 10000000.....0
- A googol is bigger than the number of atoms in the universe
- Google was founded by Larry Page and Sergey Brin, two Stanford Univ. Computer Science graduate students
- In 1998 they built a prototype system called BackRub, dropped out of school, and tried to attract investors for their new company
- Google Inc. released a beta version on Sept. 7, 1998
- www.google.com was officially released on Sept. 21, 1999



A Brief Chronology of Search Engines

Algorithmic Search Era

Gopher/Archie/
Veronica - Early
Internet Search
Engines

Yahoo

Lycos

Excite

Alta Vista

Inktomi

Hotbot

Ask Jeeves

Google

Google
begins
Adword and
Pay-Per-Click

1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003

Overture (goto.com) is the first to combine sponsored (paid) search results with conventional search results

Goto.com

introduces
pay-per-click
search results

Google

begins to include
pay-per-click
search results

Paid Search Era

(paid search became pervasive)



Search Engine Basic Behavior



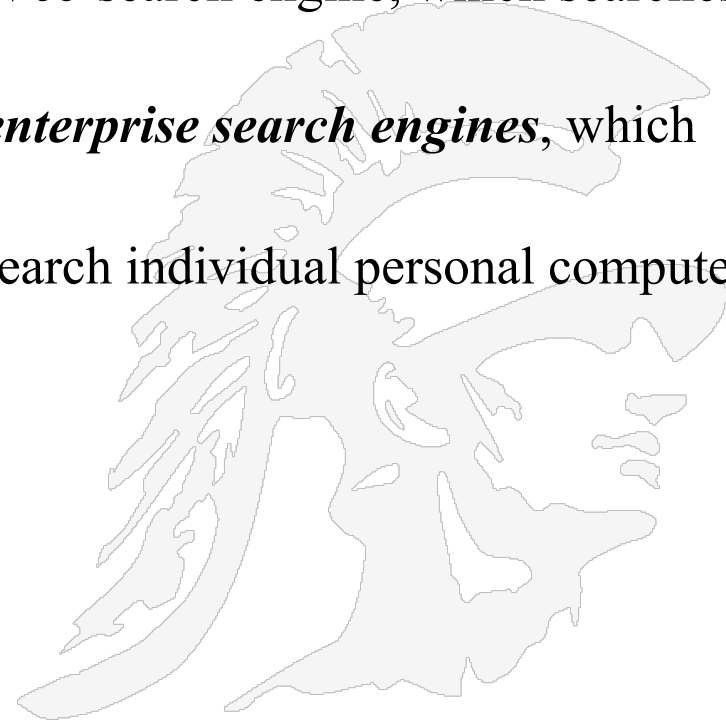


What is Web Search?

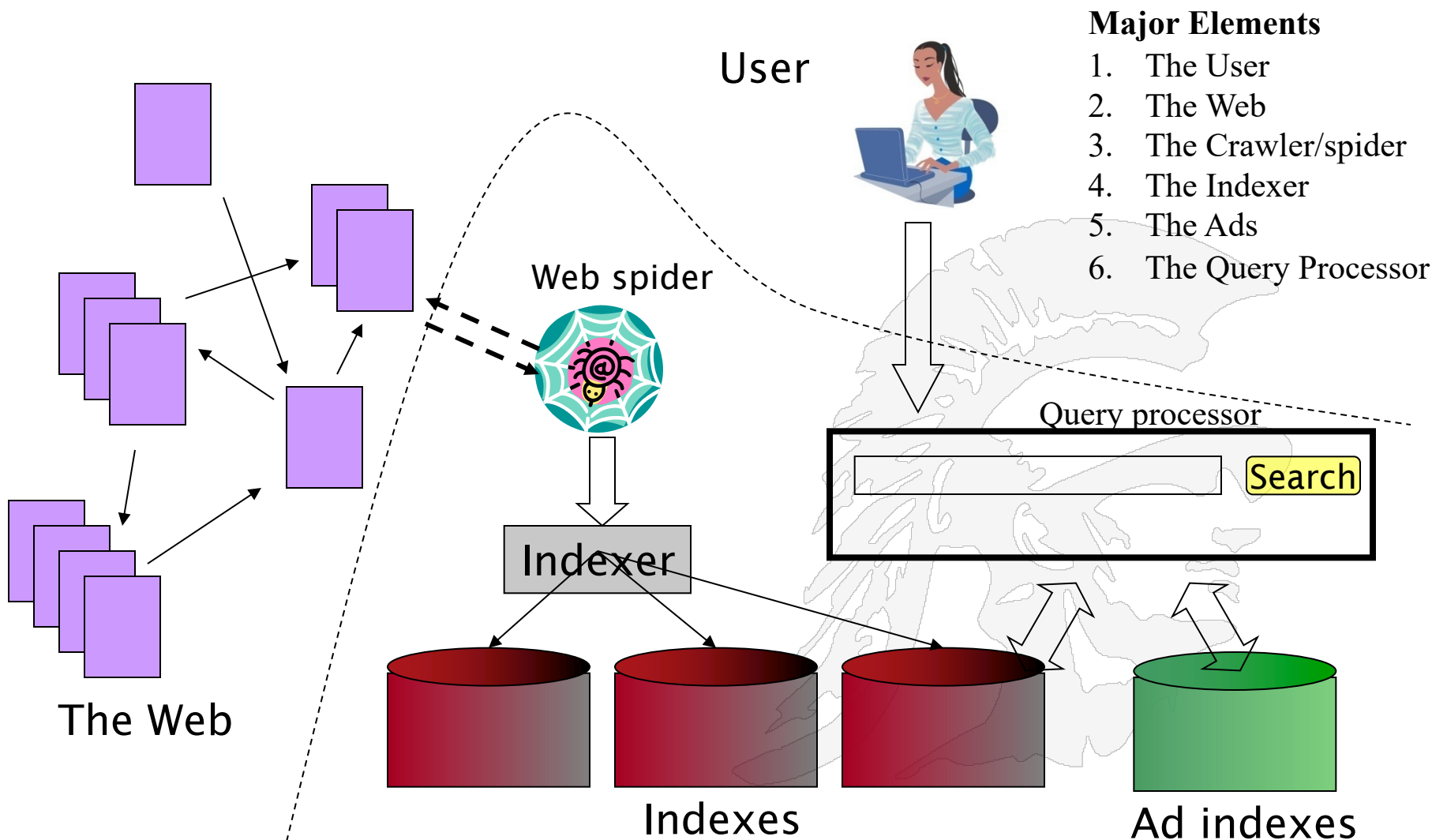
- **Providing access to heterogeneous, distributed information that is publicly available on the World Wide Web**
 - Information comes in many different formats
 - Most of the information has not been screened for accuracy
- **Multi-billion dollar business**
- **Source of new opportunities in marketing**
- **Strains the boundaries of trademark and intellectual property laws**
- **A source of unending technical challenges**

Web Search Engine Definitions

- “A search engine is a program designed to help find information stored on a computer system such as the World Wide Web, inside a corporate or proprietary network or a personal computer” *wikipedia*
 - *search engine* usually refers to a *Web* search engine, which searches for information on the public Web.
 - Other kinds of search engine are *enterprise search engines*, which search on intranets,
 - *personal search engines*, which search individual personal computers

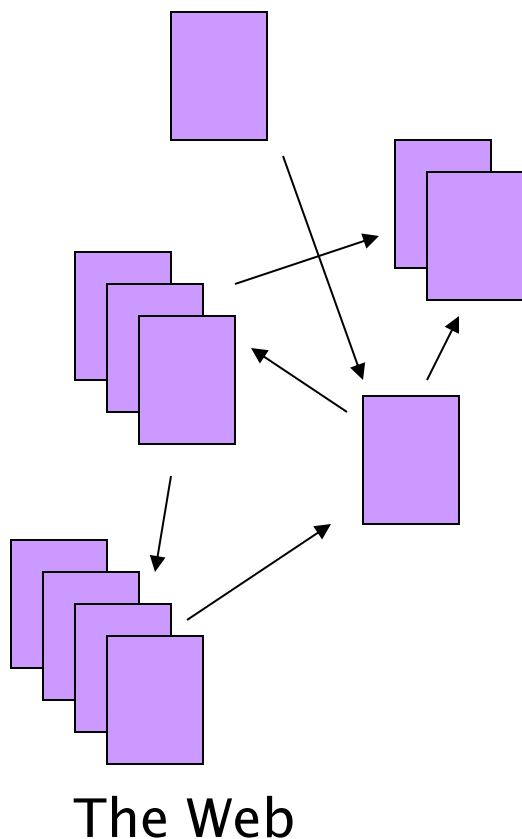


Basic Web Search Internals



Web Search Engine Elements

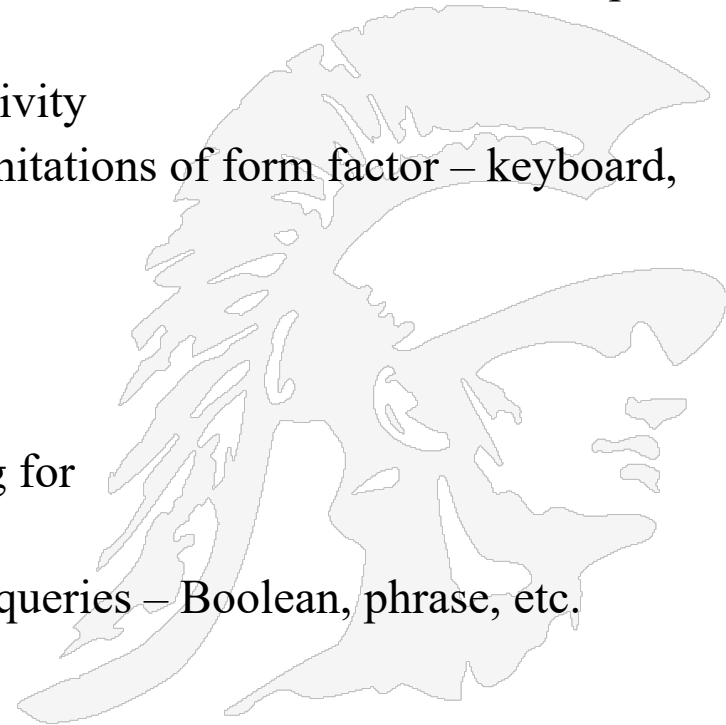
- ***Spider* (a.k.a. crawler/robot) – builds corpus**
 - **Collects web pages recursively**
 - For each known URL, fetch the page, parse it, and extract new URLs
 - Repeat
 - **Additional pages come from direct submissions & other sources**
- **The *indexer* – creates inverted indexes**
 - Various policies wrt which words are indexed, capitalization, support for Unicode, stemming, support for phrases, etc.
- ***Query processor* – serves query results**
 - **Front end** – query reformulation, word stemming, capitalization, optimization of Booleans, etc.
 - **Back end** – finds matching documents and ranks them



- **No design/co-ordination**
- **Distributed content creation, linking**
- **Content includes truth, lies, obsolete information, contradictions ...**
- **Data is stored in structured (databases), semi-structured (tables)...**
- **Scale larger than previous text corpora**
- **Growth – still expanding**
- **Content can be *dynamically generated***



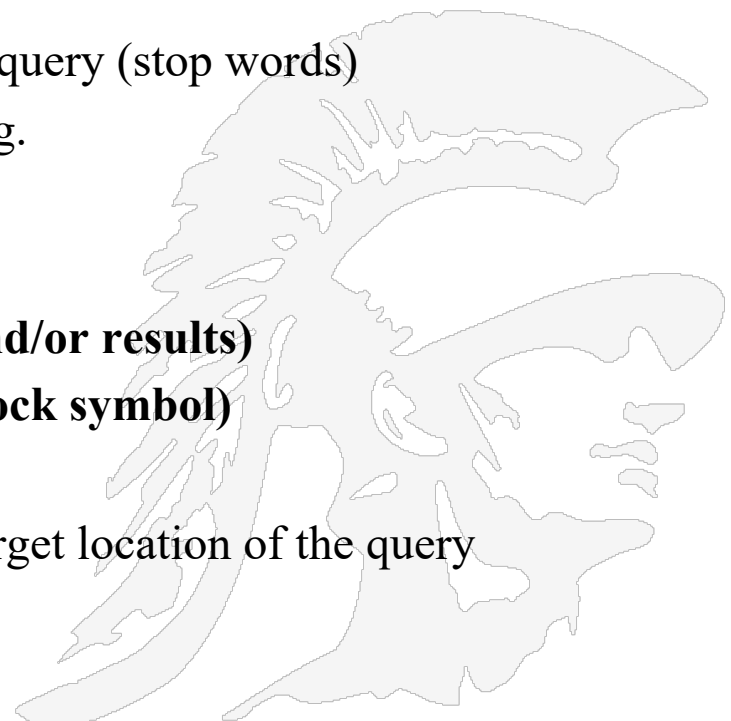
- **Diverse in background/training**
 - Users sometimes cannot tell the difference between a search bar from the URL address field (**Chrome conflates the two**)
 - Users rarely use the scroll bar, so key results must be at or near the top
- **Diverse in access methodology**
 - Increasingly, high bandwidth connectivity
 - Growing segment of mobile users: limitations of form factor – keyboard, display
- **Diverse in search methodology**
 - Search, search + browse,
 - Average query length ~ 2.5 terms
 - Has to do with what they're searching for
- **Poor comprehension of syntax**
 - Early engines offered rich syntax for queries – Boolean, phrase, etc.
 - Current engines hide these



User's Information Needs Are Diverse

- **Informational** – want to learn about something (~40%)
 - e.g. Low hemoglobin
- **Navigational** – want to go to that page (~25%)
 - e.g. United Airlines
- **Transactional** – want to do something (web-mediated) (~35%)
 - Access a service
 - Los Angeles weather
 - Downloads
 - Mars surface images
 - Shop
 - Nikon CoolPix Camera
- **Gray areas**
 - Find a good hub
 - Car rental in Finland
 - Exploratory search “see what’s there”

- Query processing involves much more than just matching query terms with document terms
- **Semantic analysis of the query includes:**
 1. Determining the language of the query
 2. Filtering of unnecessary words from the query (stop words)
 3. Looking for specific types of queries, e.g.
 - **Personalities (triggered on names)**
 - **Cities (travel info, maps)**
 - **Medical info (triggered on names and/or results)**
 - **Stock quotes, news (triggered on stock symbol)**
 - **Company info ...**
 4. Determining the user's location or the target location of the query
 5. Remembering previous queries
 6. Maintaining a user profile





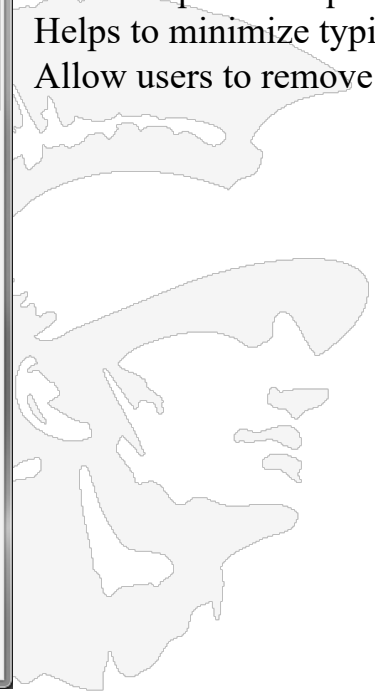
USC Viterbi
School of Engineering

Google Maintains Your Recent Query History

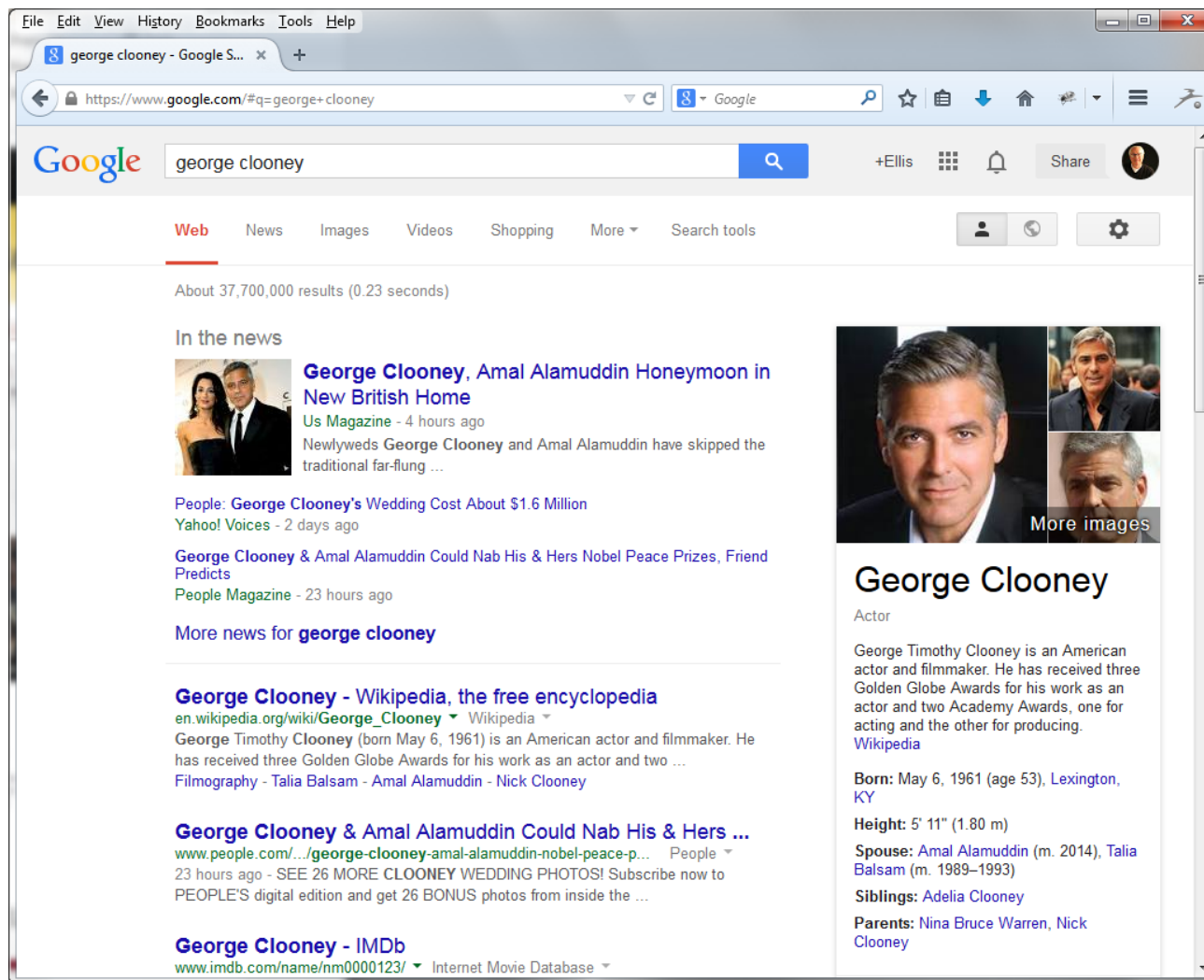
The screenshot shows a Google search interface. The search bar contains the text "weather los angeles ca". A dropdown menu is open, displaying a list of recent queries: "weather los angeles ca", "weather san francisco", "weather", and "weather san diego". Each query has a "Remove" link next to it. Below the search bar, the weather widget for Los Angeles, CA is displayed, showing a current temperature of 69°F, a clear sky, and a 7-day forecast.

Day	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu
High/Low	81° / 63°	75° / 61°	81° / 63°	88° / 64°	88° / 64°	81° / 61°	75° / 61°	77° / 61°

- Maintain previous queries
- Helps to minimize typing
- Allow users to remove old ones



Results are Holistic A Person Query



File Edit View History Bookmarks Tools Help

george clooney - Google S... x +


https://www.google.com/#q=george+clooney

Google george clooney +Ellis Share

Web News Images Videos Shopping More Search tools

About 37,700,000 results (0.23 seconds)

In the news

 **George Clooney, Amal Alamuddin Honeymoon in New British Home**
Us Magazine - 4 hours ago
Newlyweds George Clooney and Amal Alamuddin have skipped the traditional far-flung ...

People: **George Clooney's Wedding Cost About \$1.6 Million**
Yahoo! Voices - 2 days ago


George Clooney & Amal Alamuddin Could Nab His & Hers Nobel Peace Prizes, Friend Predicts
People Magazine - 23 hours ago

More news for **george clooney**

George Clooney - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/George_Clooney Wikipedia
George Timothy Clooney (born May 6, 1961) is an American actor and filmmaker. He has received three Golden Globe Awards for his work as an actor and two ...
Filmography - Talia Balsam - Amal Alamuddin - Nick Clooney

George Clooney & Amal Alamuddin Could Nab His & Hers ...
www.people.com/.../george-clooney-amal-alamuddin-nobel-peace-p... People
23 hours ago - SEE 26 MORE CLOONEY WEDDING PHOTOS! Subscribe now to PEOPLE'S digital edition and get 26 BONUS photos from inside the ...

George Clooney - IMDb
www.imdb.com/name/nm0000123/ Internet Movie Database

 More images

George Clooney
Actor

George Timothy Clooney is an American actor and filmmaker. He has received three Golden Globe Awards for his work as an actor and two Academy Awards, one for acting and the other for producing.
Wikipedia

Born: May 6, 1961 (age 53), Lexington, KY

Height: 5' 11" (1.80 m)

Spouse: Amal Alamuddin (m. 2014), Talia Balsam (m. 1989–1993)

Siblings: Adelia Clooney

Parents: Nina Bruce Warren, Nick Clooney

Includes the following:

Latest news

Biography

Photos

Basic facts

born

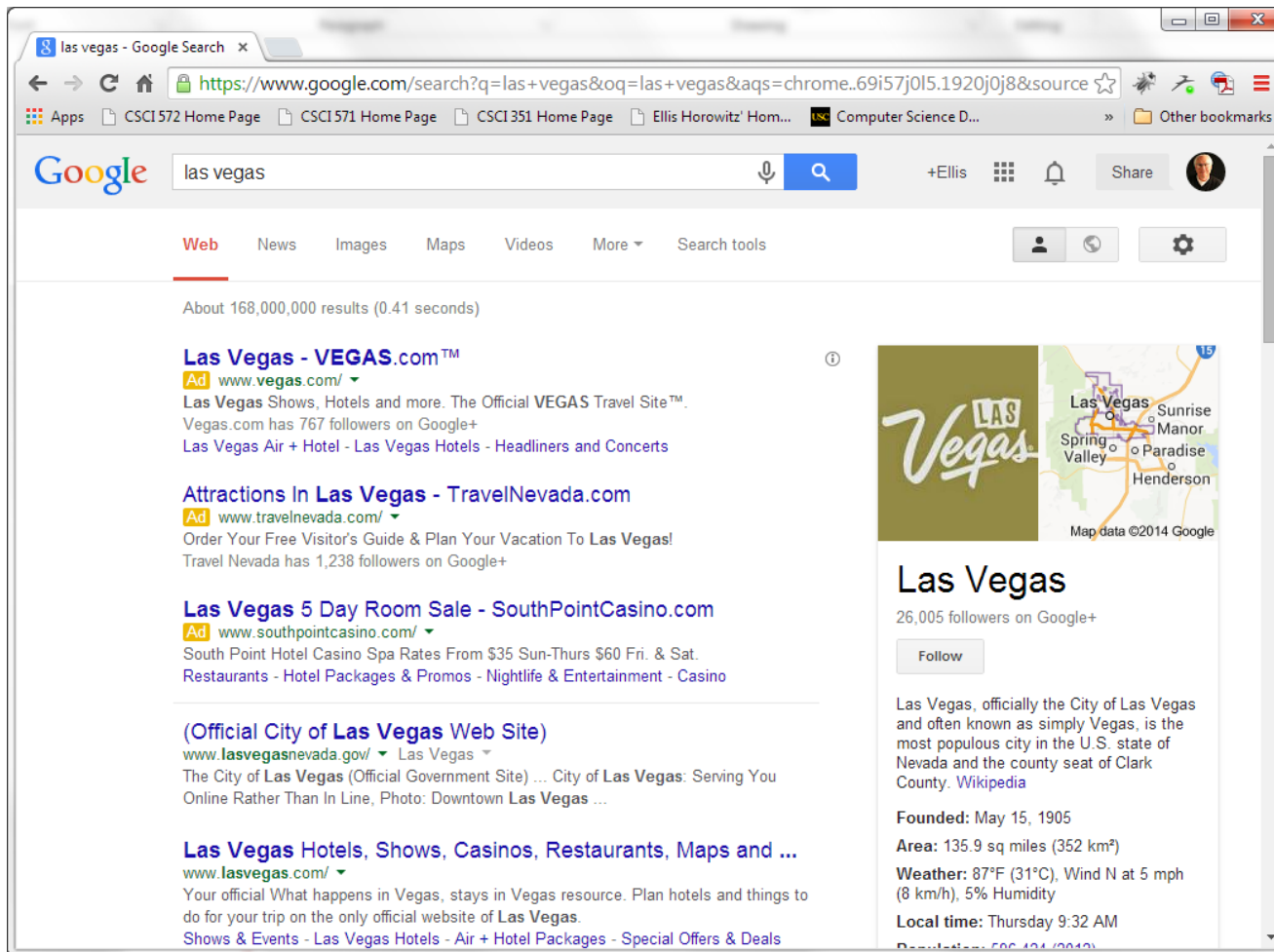
married

parents

career

Results are Holistic

A Place Query



las vegas - Google Search

https://www.google.com/search?q=las+vegas&oq=las+vegas&aqs=chrome..69i57j0l5.1920j0j8&source

Apps CSCI 572 Home Page CSCI 571 Home Page CSCI 351 Home Page Ellis Horowitz' Hom... Computer Science D... Other bookmarks

Google las vegas

+Ellis [Grid] [Bell] Share [Profile]

Web News Images Maps Videos More Search tools

About 168,000,000 results (0.41 seconds)

Las Vegas - VEGAS.com™
 Ad www.vegas.com/
 Las Vegas Shows, Hotels and more. The Official VEGAS Travel Site™. Vegas.com has 767 followers on Google+
 Las Vegas Air + Hotel - Las Vegas Hotels - Headliners and Concerts

Attractions In Las Vegas - TravelNevada.com
 Ad www.travelnevada.com/
 Order Your Free Visitor's Guide & Plan Your Vacation To Las Vegas! Travel Nevada has 1,238 followers on Google+

Las Vegas 5 Day Room Sale - SouthPointCasino.com
 Ad www.southpointcasino.com/
 South Point Hotel Casino Spa Rates From \$35 Sun-Thurs \$60 Fri. & Sat. Restaurants - Hotel Packages & Promos - Nightlife & Entertainment - Casino

(Official City of Las Vegas Web Site)
 www.lasvegasnevada.gov/ Las Vegas
 The City of Las Vegas (Official Government Site) ... City of Las Vegas: Serving You Online Rather Than In Line, Photo: Downtown Las Vegas ...

Las Vegas Hotels, Shows, Casinos, Restaurants, Maps and ...
 www.lasvegas.com/
 Your official What happens in Vegas, stays in Vegas resource. Plan hotels and things to do for your trip on the only official website of Las Vegas.
 Shows & Events - Las Vegas Hotels - Air + Hotel Packages - Special Offers & Deals

Las Vegas
 26,005 followers on Google+
 Follow

Las Vegas, officially the City of Las Vegas and often known as simply Vegas, is the most populous city in the U.S. state of Nevada and the county seat of Clark County. [Wikipedia](#)

Founded: May 15, 1905
Area: 135.9 sq miles (352 km²)
Weather: 87°F (31°C), Wind N at 5 mph (8 km/h), 5% Humidity
Local time: Thursday 9:32 AM
 Population: 596,494 (2013)

Map data ©2014 Google

Includes the following:

Official site

Map

Essential facts

founded

area

weather

time

population



Results are Holistic An Hotel Query

The screenshot shows a Google search for "sheraton times square hotel nyc". The search results include:

- Sheraton™ New York Hotel - Official Site - Our Best Rates** (Ad) with a link to www.sheraton.com/TimesSquare. It includes a "Guaranteed. Book Now!" badge, mentions 1,593 followers on Google+, and provides the address: 811 7th Avenue, New York, NY. Links for "Photos", "Make a Reservation", "Features & Amenities", and "Special Offers" are visible.
- NYC Times Square Hotel - softel-new-york.com** (Ad) with a link to www.softel-new-york.com/. It describes the hotel's location in Manhattan near Times Square and offers a direct booking link.
- Sheraton New York Times Square Hotel: Hotel Near Time...** (www.sheratonnewyork.com/). This result includes a description of the hotel's location, a 3.2-star rating from 141 Google reviews, and a price of \$239.
- Sheraton New York Times Square Hotel - Starwood Hotel...** (www.starwoodhotels.com/sherato...). This result includes a 3.7-star rating from 1,851 reviews.

On the right side of the search results, there is a map showing the location of the hotel at 811 7th Avenue. Below the map is a card for "Sheraton New York Times Square Hotel" with a "\$239 Book" button and a "Directions" button. At the bottom of the map area, there is an "Ads" section with another listing for "Sheraton Times Square Hotel" with a link to sheratonnewyork.reservations.com/.

- Includes the following:**
- Main hotel website
 - Map
 - Address
 - Phone number
 - Price of a room
 - Directions



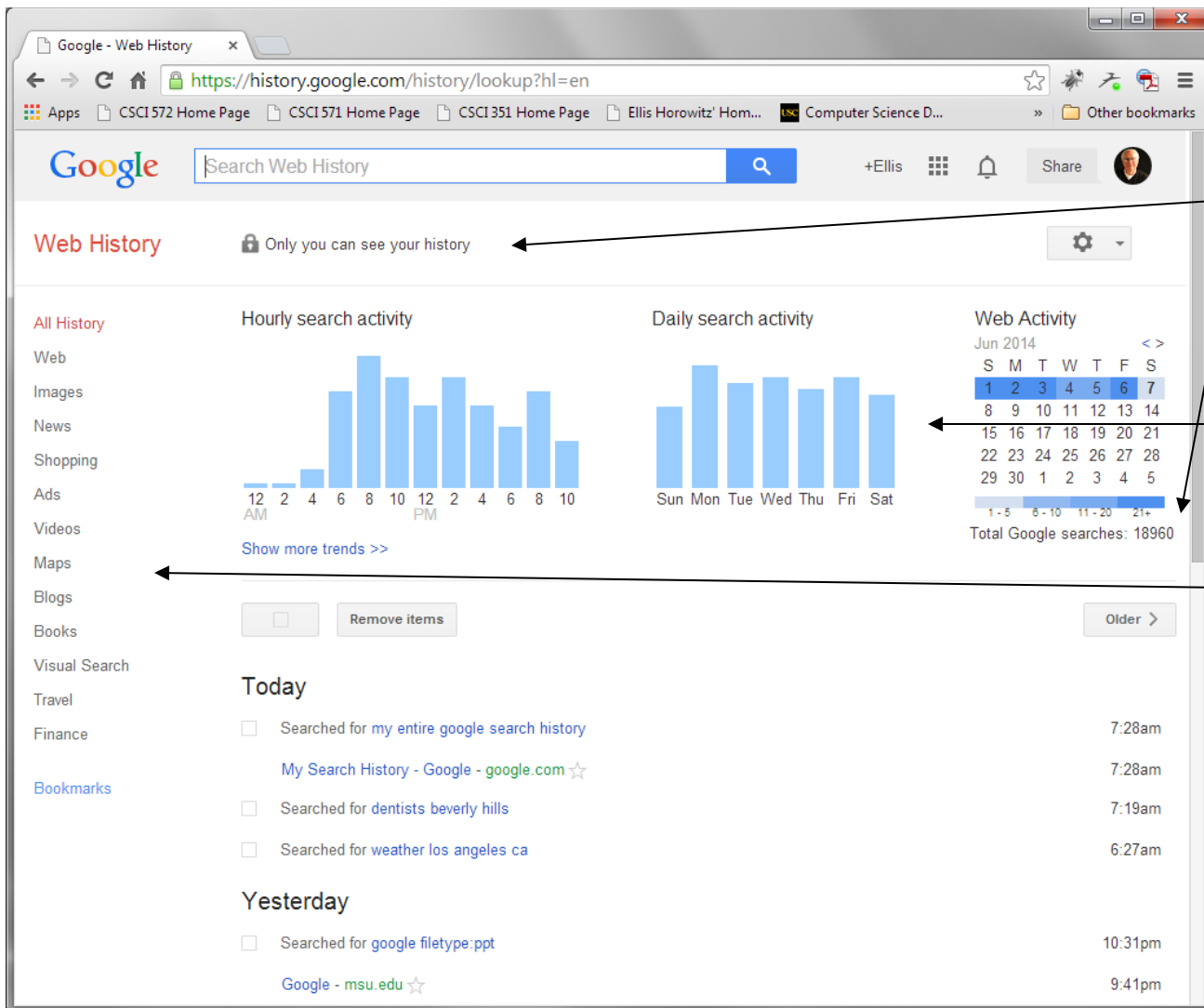
Google Retains a User's Entire Query History!

They claim that only I can see my history;
I have issued a total of 18,960 queries;

Graphs show my queries by hour and by week;

I can view my Web queries as distinct from my Image queries or my News queries, etc

As a result, Google now knows a great deal about us!



The screenshot shows the Google Web History page for a user named Ellis. The page includes a search bar, a sidebar with navigation options (Web, Images, News, etc.), and a main content area with three graphs: 'Hourly search activity', 'Daily search activity', and 'Web Activity'. The 'Web Activity' graph shows a calendar for June 2014 with a bar indicating 18,960 total searches. Below the graphs is a list of search queries for 'Today' and 'Yesterday'.

Time	Search Query
7:28am	Searched for my entire google search history
7:28am	My Search History - Google - google.com
7:19am	Searched for dentists beverly hills
6:27am	Searched for weather los angeles ca
10:31pm	Searched for google filetype:ppt
9:41pm	Google - msu.edu

Search Engines are an Industry

- **The search engine industry is 20+ years old, having started with WebCrawler and Lycos in 1994 who sold banner ads as their business model**
- **Search engine revenue today**
 - **Google:** 2021:\$257 Billion; 2020: \$181 Billion; 2019: \$162 Billion; 2018: \$116 Billion; 2017: \$109 Billion; 2016: \$90 Billion; 2015: \$74.5 Billion; 2014: \$66 Billion; 2013: \$37 Billion
 - **Baidu:** 2021: \$31 Billion; 2020: \$16.4 Billion; 2019: \$15 Billion; : \$11.3 Billion; 2017: \$13 Billion; 2016: \$10.1 Billion; 2015: \$10.2 Billion; 2014: 8.0 Billion
 - **Yahoo:** 2021: 5.2Billion; 2019: 6.97Billion; 2018: 3.03 Billion; 2017: 3.0 Billion; 2016: 2.98 Billion;2015: \$4.9 Billion; 2014: 4.6 Billion; 2013: 4.6Billion
 - **Bing:** 2020 \$7.74 Billion; 2019: \$7.63 Billion; 2018: \$7.01 Billion
 - Microsoft says that in Q1 2016 Bing became profitable



Google is a Monopoly Gatekeeper for the Internet

- **The US is suing Google for anti-Trust violations**
 - **Google claims it has strong competition in search!!!**
 - **Google has sweetheart deals with Apple and pays Apple \$8+ Billion/year to be their default search engine**
- **Google maintains the largest index of the web**
 - **Some websites actually deny access to crawlers other than Google and Bing as these other crawlers bring in little traffic and consume server cycles**
 - **Only Google and Bing have the resources to maintain such a large index, e.g. DuckDuckGo no longer crawls the web and uses Bing's index**